Appendix C.—Assessment of Medical Technology: Methodological Considerations

by Paul M. Wortman, Ph. D., University of Michigan
and
Leonard Saxe, Ph. D., Boston University*

Abstract

This appendix is primarily concerned with methodological issues underlying the research evidence used to assess medical innovations. In particular, it examines the process of research analysis in interpreting the results from individual studies and the complementary process of research synthesis in aggregating the results from many studies. Both processes are important to medical technology assessment and require an understanding of their methodological limitations. A conceptual framework is presented for determining the validity of the research evidence derived from various methodologies (e.g., clinical trials, consensus exercises) employed to assess medical technology.

Introduction

Medical technology has assumed an increasingly central role in the delivery and costs of health services. In order to assess the effectiveness of medical technologies and increase the impact of Federal funds, Congress has undertaken a number of policy initiatives over the past few years (310). Through the 1976 Medical Device Amendments, it expanded the Food and Drug Administration's (FDA's) role in assessing medical products for safety and effectiveness. In 1978, it established the National Center for Health Care Technology (NCHCT) with a mandate to conduct medical technology assessments. Technology assessment has been defined as a "comprehensive form of policy research that examines the . . . social consequences of technology" **(7,269).** Technology assessments must consider a wide range of outcomes of **a** technology, including safety, efficacy, cost effectiveness, and social impact. These outcomes are judged by considering various forms of information about a technology. This information is typically derived from multiple studies that vary in their methodological adequacy and appropriateness to assess the technology.

● **Authors'** *acknowledgments:* The authors would like to thank Drs. Fred Bryant, John McSweeny, William Yeaton, and the many anonymous reviewers for their helpful comments on previous drafts. The authors also thank Marga Van Goethem and Jean Holther for the many hours of conscientious typing.

The authors accept full responsibility for any remaining infelicities of prose and inaccuracies of thought. The views expressed in this paper are those of the authors and do not necessarily represent the views of OTA.

Despite their importance, methodological features of the research studies used to develop a technology assessment are often given only minimal attention. Methods that have very different functions and applicability, such as controlled clinical trials and consensus development, are often lumped together and viewed as alternatives to one another (see **266).** Similarly, randomized clinical trials (RCTs) are often seen as a unitary method, although they represent a diverse set of procedures. In addition, while the usefulness of a technology assessment rests on the ability to integrate research evidence, little attention is paid to research synthesis activities. There are no clear-cut standards for the quality of evidence that should be considered nor for the ways in which discrepant information should be consolidated. Recent Government conferences on methods for assessing medical technology have not changed the situation (e.g., **3,87).**

The pressures for diffusion of medical innovations require valid statements of efficacy, safety, and social impact (266). These, in turn, necessitate appropriate methods for assessment. Proper research methods allow one to state with confidence that observed effects are actually due to the medical innovation—i.e., well-designed and carefully conducted evaluative studies for technology assessments will produce valid and reliable results. As the remainder of this appendix will demonstrate, the failure to conduct proper studies often results in serious criticism of both the validity of the research and the validity of the technology assessments **based on this** research. A framework for determining validity that can be used to interpret the results of individual studies and to synthesize the findings from many studies will be presented.

The purpose of this appendix is to review some principles for interpreting and integrating the results of evaluative studies that underlie the assessment of medical technologies and to indicate their place in a general strategy for medical technology assessments. The remainder of this appendix is organized in three sections. The section immediately below discusses the interpretation of individual evaluative studies of medical technology, It introduces validity concepts and describes the relationship between the design of research studies and the usefulness of the information generated. Three broad categories of designs are discussed: 1) RCTs; 2) controlled clinical trials lacking randomization, also known as quasi-experiments (45);

and 3) uncontrolled studies known as nonexperimental investigations or case studies. Some typical designs are described, and the problems they pose in interpreting the evidence from studies assessing health care technology are presented. The second section below examines methods for synthesizing the results from many studies. These include formal quantitative procedures (e.g., meta-analysis) and group decisionmaking techniques (e.g., consensus conferences). The validity problems in using these procedures are discussed. The final section briefly describes a strategy for integrating these assessment methods with the innovation process.

Research Analysis: Interpreting the Results of Individual Studies

A thorough technology assessment is viewed as including 10 elements (269), one of the most important elements in a technology assessment is the "evaluation of potential impacts, " which encompasses "technical feasibility" (i.e., effectiveness), safety, ethics, and economic considerations. If technology assessments are to be useful, their evaluation component must be conducted in a systematic manner that employs acceptable scientific methods, especially research design (60,144). Proper research design is of utmost importance if the observed changes in a patient population are to be correctly attributed to the technology being assessed rather than to some extraneous factors. As the following discussion will show, it is often these other unrelated factors that cloud the interpretation of technological impact and undermine the validity of the technology assessment.

Validity

Validity involves the careful analysis of research to determine its adequacy or scientific soundness. The analysis of research requires an understanding of the strengths and weaknesses of the methods used to generate scientific evidence. The problems in determining the validity of the findings in research studies have been of continual interest in medicine. Recently, the medical journal Lancet (170,171) carried a series dealing with research design issues in "assessing clinical trials. " The articles discussed problems that can undermine the validity of clinical trials, especially those dealing with medical innovations. A useful conceptual framework for examining issues of validity has been developed by Cook and Campbell (68). These methodologists organize validity problems into four categories: 1) internal validity, 2) statistical conclusion validity, 3) external validity, and 4) construct validity. These four categories provide a useful way of understanding the implication of design issues for medical technology assessment studies.

INTERNAL VALIDITY

Internal validity refers to whether the observed effects of a medical innovation are truly due to the technology and not to some other factors. Internal validity, therefore, is the most important component of validity. An important part of any technology **as**sessment asks questions such as: Would patients have improved even if they did not receive the innovation? or, Do they really improve more with the innovative procedure than with the traditional approach? An evaluative study that can adequately answer these questions is called an internally valid evaluation. From a scientific perspective, internal validity involves the assignment of causality to the innovation for the observed benefits or risks.

A key issue in the internal validity of an assessment is the "control" of factors extraneous to the innovation. When random assignment of patients to treatment and control groups fails or is not employed, a number of plausible alternative explanations can be offered. These so-called "threats to validity" include, among others, alternative hypotheses based on selection and statistical regression (5,68). Selection or selection bias occurs when patients are assigned to receive a treatment because of particular characteristics (e.g., better prognosis), while ***statistical regression*** arises when patients are chosen because of their extreme value on a laboratory test or other measure relevant to the treatment. Many of these validity threats are defined, described, and discussed in the following sections. Their usefulness in interpreting the evidence from technology assessment studies will be demonstrated.

A properly conducted RCT is internally valid. Even when studies are advertised as RCTs, however, one should carefully examine their methods or procedures to determine if the randomization process was properly conducted. A recent RCT published in the ***Journal of the American Medical Association*** by Hoehler, et al. **(1.90),** illustrates the problem. To assess the effectiveness of a rotational spinal manipulation for back pain, the authors report, 95 subjects were admitted to the trial and were "randomly assigned to either the experimental or the control group." From this brief description of the randomization procedure, one would expect that 45 to **50** subjects would be assigned to each condition. Instead, the initial table reveals that there were **56** in the experimental, spinal manipulation condition and 39 in the control group. This ap-

pears to be quite divergent from what a randomized process would produce. (The probability of this difference occurring is greater than the "1 in **20**" level associated with chance.) Although there may be good reasons for this discrepancy besides chance, the authors are mute on this point. One is left with the suspicion that other factors (e.g., severity of pain) may have influenced patient assignment to conditions and. that these selection factors may be responsible for the observed results. These factors would pose a threat to internal validity due to differential patient selection into the two groups.

STATISTICAL CONCLUSION VALIDITY

There are many threats to the validity of a technology assessment study. Threats related to the analysis of the data are particularly important, and Cook and Campbell have called these threats to statistical conclusion validity. This category of validity focuses on the appropriateness of statistical tests and their ability (or power) to determine whether or not observed effects are due to chance. Many, otherwise internally valid, studies in health have used too few subjects (see 153,171) to detect anything but the largest effects. Statisticians call this a Type II error—the acceptance of a finding of no difference (in effectiveness) when it is false. It is possible that some useful technological innovations have been discarded due to faulty statistical procedures. For example, a recent study (264) on the effectiveness of timolol in reducing mortality after a heart attack noted that one reason most other studies of these beta-blockers have found little or no effect was that they contained too few patients "to exclude the possibility that a beneficial effect was being overlooked. "

EXTERNAL VALIDITY

External validity concerns the generalizability of the observed effects to other patient populations, settings, or conditions. That is, would the treatment be beneficial in other settings or are its effects specific to the present situation? The concept of external validity is captured in OTA's definition of "efficacy" (266), the likelihood of benefit under optimal circumstances to "individuals in a defined population . . . ." The importance of external validity considerations can be found in an example drawn from the first National Institutes of Health (NIH) (96) consensus development conference on the efficacy of mammography in the detection of breast cancer. The panel concluded that the technology was only beneficial for women over 50 and might be harmful for others due to the risks of repeated exposure to radiation. These conclusions, based largely on one study (341), indicated dramatic differences in effectiveness from one subpopulation of women to another.

CONSTRUCT VALIDITY

The last type of validity deals with conceptual issues. It depends on the adequacy of the theory that one has about what makes the innovation effective and the adequacy of the measures of the observed effects (or variables) derived from the theory. The recently concluded debate on the efficacy of radical mastectomy demonstrated the role of theory (147). Once it was shown that cancer was disseminated through the blood stream, the basis of the Halsted radical surgery was called into question. Construct validity also refers to improper measurement of outcomes as well as improper control of the technology. The latter can often be confused or contaminated by other changes that may cause the observed effects.

Outcome Measures.—One of the major problems in assessing medical technology is the absence of good outcome measures of the constructs considered important. For example, a researcher in behavioral medicine attending a conference on the social impact of coronary artery bypass graft (CABG) surgery noted, "There is no consensus about how you define and measure quality of life" (284). As a consequence, technology assessment studies often focus on a variety of process variables (e.g., admissions, length of stay, etc.) that may, or may not, be indicative of the delivery of services and are not concerned with the overall impact of a technology on patient health. Often, the absence of such observations can be traced to the lack of a specific, well-defined treatment procedure.

Even where there are outcome measures *or* end points, these may be "soft"* or subjective. Relief of angina in CABG surgery is a case in point. Both patients' and physicians' expectations concerning the benefits of surgery (see 297) may influence judgments of relief. Such expectations are the rule for the technological advances in modern medicine. As discussed, it is essential to eliminate from technology assessment studies the potential bias produced by these expectations of efficacy (i.e., placebo effects). Good measures of the impact of innovative treatments are needed, Often, the debate on the efficacy of medical technologies swirls about very few objective outcome measures (e.g., survival in CABG surgery, cesarean section in fetal monitoring).

*Paul Meier, University of Chicago, personal communication, December 1960.

Design Categories

Several types of designs have been used in studies evaluating or assessing medical technologies. These generally fall into the three categories noted above: 1) RCTs, 2) quasi-experiments or controlled trials, and 3) uncontrolled case studies. These designs represent the principal methodological approaches to research studies of medical technology assessment and vary with respect to the validity of the evidence they produce. In this section, the advantages and disadvantages of each design category are examined with respect to validity. The major concern in this discussion is with the choice of an appropriate control or comparison group and the effect this has on the validity of the findings.

RANDOMIZED CLINICAL TRIALS

The "true" experiment or RCT is the preferred design for producing unambiguous assessments of a medical technology (see 60,187). The essential ingredient in an RCT is randomization: Patients or other experimental units are randomly assigned to experimental (treatment) or control conditions. Although some **(76)** argue that only posttreatment measurement of patients is required in an RCT, most health researchers use both pretreatment and posttreatment measures. This provides a check on the initial or baseline equivalence of the groups and an accurate (or unbiased) estimate of the amount of change produced by the innovation. The basic question asked in a true experiment is whether effects observed in the experimental (or treated) group are also observed in the control (or untreated) group. If the answer is essentially "no," the effects may be safely attributed to the technology.

RCTs are in reality a family of designs that vary in size and complexity. The number of treatment conditions can vary (e.g., dosage levels) as can the size of the population and the theoretical significance of the study. Small randomized trials are often performed early in the development of a technology to demonstrate or test the efficacy of the treatment's innovative elements. Such studies typically involve only a single investigator observing a few subjects—either animals or humans—at a single site. At another level, large-scale, multicenter trials are often conducted to establish the efficacy or safety of a developed technology. Such RCTs are usually necessary to provide the appropriate number and type of patients to assess the technology as quickly as possible. Moreover, the diversity of sites and subjects can provide useful data on the external validity of the innovation, These multicenter trials are not immune from problems. They add difficulties in organizational complexity and hence limit the researcher's ability to assess the technology under "ideal conditions." Indeed, much of the debate over the Veterans Administration's (VA's) multicenter RCT of CABG surgery centered on such problems **(254,255),** There were wide differences both in types of patients selected and in the operative mortality among the sites. Comer (66) discusses several strategies for successfull maintaining the integrity of the randomization process (e.g., a centralized procedure with few implementers).

Blinding.—Other attributes of RCTs are also important to note. In order to reduce the bias in physician and patient expectations, physicians and patients should both be unaware of or "blind" to the treatment the patient is receiving. This is called a "double-blind study" and is frequently used in assessing drugs. In some cases, such control is not possible. For example, today it would be ethically impossible to give some patients sham surgery to assess the efficacy of CABG surgery or even to give them the much simpler internal mammary artery ligation surgery that was proven ineffective using such a control group (20). And even if it were possible, only the patients would not know which "treatment" they had received (i.e., the study would be a single-blind study). As noted above, the inability to blind patients and physicians contributes to construct validity problems in interpreting the surgery's effect on the relief of angina.

When researchers and patients are not blind to the treatment being delivered, it is possible that their expectations can affect (or be confounded with) the outcomes. To avoid this, RCTs often use a placebo (i. e., a procedure that appears identical to the innovation but has no therapeutic benefit). A good example of a relatively uncomplicated RCT employing a placebo is provided by The Coronary Drug Project (72) that assessed the effectiveness of a drug using this technique. Placebo control is useful in establishing construct validity but is hard to employ with most non-drug innovations.

The Hoehler, et al. (190), study of spinal manipulation for relief of back pain (noted above) is an exception in that it employed a placebo treatment for an assessment of a "technique. " The control group patients received a "soft-tissue massage of the lumbrosacral areas." The authors assumed that this was a valid placebo because their previous research showed that "patients with no knowledge of spinal manipulation probably cannot distinguish that therapy from soft-tissue massage. " They found no significant difference at discharge as both groups were substantially improved. In fact, the observed "dramatic" effects of a number of innovations (e.g., gastric freezing and internal mammary artery ligation) were later shown by well-designed RCTs to be due to a "placebo effect."

Chalmers, et al. (54), maintain that research investigators must also be blind to the randomization process and to the interim results while the trial is in progress. In the former situation, bias could affect patient assignment; in the latter, it could also affect patient withdrawals. In either case, the validity of the study is jeopardized.

One should be cautious, however, in assuming that all RCTs are necessarily exemplary and immune to threats to their validity. Research reports often conceal major flaws in the conduct of the RCT. The recent critique of the Anturane Reinfarction Trial by FDA (363) provides a cogent illustration of the problems that can occur. The FDA audit found major errors in coding outcomes and classifying patients that undermine the credibility of the study. For example, errors made in the assignment of cause of death systematically favored finding a benefit for sulfinpyrazone (Anturane) in reducing mortality following myocardial infarction. Moreover, the classification scheme itself was found to be lacking in meaning (i.e., in construct validity).

Statistical conclusion validity is also important in assessing RCTs. Most of the RCTs on coronary bypass surgery have data analysis problems stemming from serious attrition (or ***experimental mortality)*** in the medically treated condition, with patients crossing over into the surgery group. The various analytic approaches for handling this problem have been inadequate (400), including those based on initial patient assignment or "intention-to-treat" **(289).** A reexamination of the crossover problem indicates that such inappropriate statistical analyses may result in a Type II error. If the worst medical cases are switching (as is indicated), then the mean outcome for their group is being inflated. For example, a simple algebraic calculation indicates that the observed amount of crossover (i. e., one-sixth) by the worst medical patients would increase the mean by at least one-fifth of a standard deviation (i. e., 0.2 SD) or 20 percent. Since survival data usually have a negatively skewed distribution, the increase in the mean could be more. Thus, it is possible that the crossovers in the coronary bypass RCTs conceal a surgically significant difference larger than **25** percent between the two groups.

Conclusions.—Although it is often true that large-scale randomized experiments are more expensive to conduct and require more planning than nonexperimental designs (see 222), that is not always the case and they should not be rejected out of hand. Reviews of health research practices indicate that the use of inexpensive nonrandomized designs often produces costly errors, since faulty results can lead to incorrect conclusions and inappropriate policy decisions (42,158, 335). Gastric freezing provides a classic example (143). Hundreds of devices were purchased by physicians based on the evidence from poorly designed, nonrandomized studies using few patients. In many cases, the greater confidence in the results of an assessment that an RCT permits greatly outweighs any difficulties in its implementation.

CONTROLLED CLINICAL TRIALS

Despite the advantages of randomized experiments, they are often difficult to implement in settings such as hospital clinics and physicians' offices. McKinlay[2] has pointed out that RCTs are especially difficult to conduct for existing technologies that are already widely diffused. Unfortunately, widespread diffusion has been a frequent occurrence in assessing medical innovations (see **253),** In such situations, administrators are usually reluctant to make the changes in policies and procedures needed to conduct a randomized experiment. Another important obstacle to conducting RCTs that is common in health evaluations is the a priori conviction of medical personnel that specific patients are best suited for the innovative treatment being evaluated. In this case, staff will resist and possibly even subvert the randomization process. For example, the assessment of high-oxygen environments as a cause of retrolental fibroplasia in premature infants was impeded by well-intentioned nurses (346). In one study, nurses raised the oxygen level for the experimental group babies in the belief that the low-oxygen environments were harmful. In another study, it was necessary to implement the treatment only partially, until evidence of the harmful effects of oxygen were more apparent. The corruptive behaviors derived from preconceived attitudes pose an additional barrier to conducting an assessment study in an applied field setting.

Sometimes researchers find that conditions prohibit RCTs. This problem can occur for a variety of reasons: politics, as noted below in the Salk Vaccine Trial; corruption of the design through attrition or other implementation problems; ethical prohibitions where patients or physicians have been persuaded of the efficacy of a treatment (see 171); or cost considerations where funds for a long-term local study are unavailable, Sometimes, unfortunately, nonrandomized studies are conducted because of a naive belief in the ability of statistical techniques to correct for the biases introduced by selection.

When randomized experiments are not feasible, investigators often use one of several quasi-experimental

[2] J. B. McKinlay, "From 'Promising Report' to 'Standard Procedure': Seven Stages in the Career of a Medical Innovation," *Milbank Mem. Fund Q.* 59:374, 1981.

designs (see 45). Quasi-experiments involve the use of self-selection procedures in the assignment of patients to either treatment or control conditions. These designs do not permit the rigorous controls provided by RCTs. Even good quasi-experiments allow some competing explanations for observed treatment effects. In particular, two quasi-experimental designs—the cohort design and the time-series design—are commonly used. The validity of these designs is discussed below.

Cohort Design.—The cohort study or nonequivalent control group design (NECGD) is the quasi-experiment that results when random assignment of subjects to the treatment and control conditions is not employed (see table C-1). Because random assignment is not used, the "treatment" and "control" groups are "nonequivalent" and may differ in systematic ways. In the discussion that follows, the term "comparison group" is used instead of "control group" when that situation obtains.

Roos, et al. (319), employed a cohort or NECGD to determine the effectiveness of tonsillectomy with or without adenoidectomy. Using claims and patient registration data provided by the Manitoba Health Services Commission, the investigators were able to create two comparison groups to assess the impact of these surgeries on subsequent episodes of respiratory illness. The first, and larger, group consisted of operated and nonoperated persons under the age of 14 covered during a 3-year period, whose records indicated evidence of tonsillar illness. For the experimental (operated) group there had to be data available for 1 year before and 1 year after their surgery. The records of the comparison group had to indicate that they remained unoperated during this period.

A number of threats to the internal validity of this study were examined by Roos, et al. (319). Since both treatment and comparison groups were similar in age and sex, it was felt that ***maturation*** (i. e., changes in health with age) was not a threat to validity. Moreover, by using concurrent controls, ***history*** (i.e., the effect of temporal events such as new health practices) was also eliminated as a threat. However, ***'local' history (68)*** (i.e., dealing with familial or physician factors such as predisposition toward surgery) may have differed among the two groups. To reduce the influence of this potential threat, a second comparison group, composed of the siblings of those operated on, was also used.

Table C-1.—Nonequivalent Control Group or Cohort Design

	Pretest		Posttest
Treatment group	0	X	0
– R			
Comparison group	0		0

0 = Observation or measurement
X = The application of the treatment or technology
– R = Absence of randomization

SOURCE: Adapted from T.D. Cook, and D. T. Campbell, Quasi-Experimentation: Design *and Analysis of Research in Field Setting,* 1979.

Statistical conclusion validity (i.e., the correctness of the data analysis) was also strengthened by using two different analytic approaches as well as subsidiary analyses to eliminate effects due to statistical regression. This latter threat was examined by stratifying the two groups according to the number of preoperative episodes of respiratory illness. If regression was causing or influencing the results, there would be greater changes in the persons with the most preoperative episodes (i.e., the extreme scores). In all cases, the results were the same—i.e., the operated group showed (statistically significant) fewer postoperative cases of respiratory illness. Additional analyses to control for the severity of the illness (by examining specific diagnostic categories) yielded similar results.

The findings of this study are not meant to be definitive with respect to the effectiveness of tonsillectomy. This procedure has become the focus of some debate with the advent of antibiotics **(389),** and RCTs are currently being conducted. However, the study is an instructive methodological example that illustrates the assessment of actual practice or the "effectiveness" of an innovation.

Matching.—One common form of the cohort design involves examining naturally occurring patient populations (as in the above example on tonsillectomy) to determine whether they differ on important characteristics and then statistically adjusting for these differences. These procedures are referred to as matching or retrospective matching. Two examples from the literature on CABG surgery illustrate the approach and its problems.

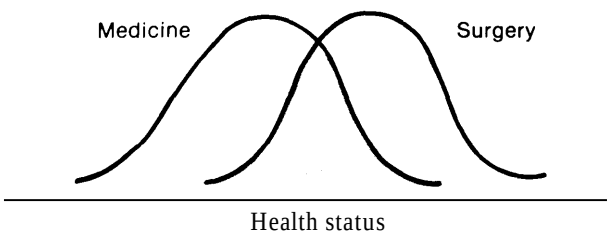
McNeer and his associates ***(240)*** examined the data drawn from 781 consecutive patients treated for coronary artery disease at Duke University Medical Center between ***1969*** and ***1973.*** Of these patients, ***402*** were treated medically and 379 had bypass surgery. Patients were compared on 89 baseline variables. The authors believed that "therapeutic decisions tend to be random." They found the two groups to be "remarkably similar" and the results to be unchanged when individual variables were corrected or statistically adjusted for initial differences. However, as Ross ***(326)*** noted in his review of this study, there was a systematic pattern of differences among significant variables such that the "surgical cohort would have a better prognosis irrespective of the form of therapy . . . ." For example, surgical patients had (statistically significant) more positive exercise tests, higher ejection fraction, and smaller heart size. The separate analyses and

adjustments do not correct for this systematic bias, and the conclusions are therefore rendered suspect. It is possible that the outcomes merely reflected the preexisting differences among the two groups.

One common form of this design involves the direct matching of patients receiving different therapies or treatment. A recent example of this method is the study conducted by Hammermeister, De Rouen, and Dodge (181) to assess the efficacy of CABG. Data from the Seattle Heart Watch angiography registry were used to form 287 matched pairs of surgical and medical patients. The patients were matched on seven variables (e.g., ejection fraction, arrhythmia, number of stenotic arteries). An analysis of the actuarial survival rates resulted in a statistically significant finding indicating decreased mortality for patients treated surgically. When the data were analyzed by the amount of coronary disease (i.e., one-, two-, or three-vessel disease), improved survival due to surgery was detected only in the subgroup of 97 pairs with two-vessel disease.

Campbell and his associates **(43,44)** have graphically demonstrated the problems posed by a matching design. In particular, they note that ***statistical regression*** to the mean (usually abbreviated as "regression") is a major threat to the internal validity of the results in such designs. The basic regression phenomenon can be easily illustrated. Assuming that the two groups (in this case, medical and surgical patients) differ on some relevant unmeasured variables, it is possible that they may be drawn from populations that differ in their health status (see fig. C-1). Given that surgeons are likely to select the best candidates for this procedure, the assumption seems warranted. The resulting matching procedure would then pair medical patients above their group's mean with surgical patients below their group's mean. Given the imperfect (or unreliable) measures used, the two groups will regress to their respective means due to this statistical artifact. The reason is that the extreme scores of the matched patients also include an extreme "score" on the "error component" or unreliable part of the measure representing the many unmeasured variables. By chance alone, this unreliable component will be less extreme the next time the measure is taken. This can cause or contribute to the finding of a statistically significant difference as the two groups regress to different means.

Figure C-1.—**An Example of Statistical Regression Resulting From Matching**



The report on the Duke registry by McNeer, et al. ***(240),*** clearly fits this picture. The surgical patients in that study were drawn from a "healthier" population, As Hammermeister, et al. (181), acknowledged, "there are probably additional unmeasured or undescribed variables of prognostic significance" in such data. Although these investigators are skeptical that this can alter the results, accumulated evidence indicates that regression can produce spurious statistical findings. There are no foolproof statistical remedies to this problem, but there are some recently developed analytic techniques that can partially adjust for measurement error (206). These approaches may be useful in situations where there are multiple measures of health status and a conceptual model specifying the presumed relationships among the variables involved. This technique would improve the statistical conclusion validity problems associated with this design.

The problems in matching indicate the difficulty in overcoming differences resulting from selection in a nonrandomized study design. The inability of statistical techniques to remove or adjust away these differences is graphically illustrated by the results of a recently reported study of the effects of drugs on coronary heart disease (73). Significant differences were found in the 5-year mortality rate for adherers (15 percent) and nonadherers (28 percent) in the placebo control group. A multivariate statistical analysis employing 40 baseline variables was performed to adjust for the differences in adherence. The adjusted mortality rates were only 16.4 and ***25.8*** percent, respectively. The baseline characteristics accounted for only a small amount of the initial difference. The authors noted that there must be unmeasured variables such as alcohol consumption and personality characteristics that can account for this difference.

Retrospective Case-Control Study .—Perhaps the most difficult variant of the cohort design is found in the field of epidemiology where retrospective case-control studies are frequently used to establish causal processes. This design consists of a group of people with a disease (i.e., the cases) who are compared with another group without the disease (i. e., the controls) to determine if they differ in their exposure to a presumed causal agent. The major problem in this design is in the selection of the comparison (or control) group. This is the major threat to the validity of this family of designs, because it is not possible to adjust for initial differences or to ensure that the treatment and

control groups are equivalent. The problem faced by the epidemiologist-researcher is considerable, because the retrospective nature of the design implies no control of the treatment.

A recent dispute over the role of estrogen therapy for postmenopausal women as a cause of endometrial cancer illustrates the problems encountered in using this research design to assess technologies. The major point of contention among researchers (191,196) concerned the appropriateness of the control group. The traditional approach in this area had been to select women with other forms of gynecological cancer. Studies using this selection procedure have found a consistently high association between endometrial cancer and estrogen use.

This method of selecting controls has been criticized for not correcting a bias among the target cases that favors the obtained result. Specifically, it has been claimed that estrogen is associated with uterine bleeding and that this condition normally leads to careful overrepresenting in the population of confirmed endometrial cancer patients. To counteract this potential selection bias in choosing cases, Horwitz and Feinstein (191) recommend the use of women being treated for uterine diseases by either dilation and curettage or hysterectomy. These women, they argue, will include many referred because of vaginal bleeding. The use of such a population to create both treatment cases and controls will adjust for the bias resulting from increased surveillance and detection. Using both selection procedures, Horwitz and Feinstein demonstrated a reduction in the likelihood of estrogen causing cancer from about 11 to a factor of about **2.**

Critics of this alternative selection approach claim that there is little or no detection bias since most cases of endometrial cancer are eventually diagnosed **(196).** They maintain that the alternative controls used by Horwitz and Feinstein are biased because they exhibit many benign conditions not normally detected. Moreover, estrogen may cause some of these other uterine diseases. Consequently, estrogen would be overrepresented in the controls. As Cole ***(61)*** has stated, patients undergoing the same diagnostic procedure as the cases can be "an inappropriate control group" since the same causal agent may be responsible for their illnesses.

In conclusion, it is important to note that the results generated by this design are essentially correlational and do not lead to unequivocal causal inferences. Horwitz and Feinstein[3] located 17 medical "topics" where multiple case-control studies reached differing conclusions, Selection bias (i.e., "avoidance of constrained controls") was the most frequent methodologic problem involved in the **17** disputes. The two approaches for constructing a control group discussed above can be viewed as providing a range of estimates for the relationship being examined. Because of the internal validity problems associated with this design, the use of different control groups to bracket the range of relative risk estimates should be considered. This would also improve construct validity in those instances where the effects of the technology are not well understood. Multiple case-control studies can also play a useful role in generating or confirming candidates (or potential causes) for unanticipated negative findings (e.g., toxic shock syndrome). In these instances, this epidemiologic approach is on the methodologic frontline of medical technology assessment. Often, where the event is rare and the number of cases is small, it is the only available method for making an assessment —e.g., of the role of aspirin in Reye's Syndrome. As with the Horwitz and Feinstein critiques, multiple studies using different controls were necessary before the association of aspirin to the disease was considered established.

Historical Controls.—Innovations often diffuse so rapidly and completely that the potential for untreated controls is greatly reduced or eliminated (see 253,266). In such a situation, researchers typically are forced to use a variant of the NECGD or cohort design that employs historical control groups—i.e., patients treated prior to the innovation. The important change in the design is a temporal one; patients in the comparison group are no longer treated concurrently with the experimental group, Some problems with the historical control group design are illustrated by a recent article discussing the use of adjuvant chemotherapy for treating osteogenic sarcoma **(215).**

Following the development of this treatment in the early 1970's, researchers began to experiment with ways to improve its apparent effectiveness. One approach was to treat patients with the drugs before their cancer had metastasized. Historical controls drawn from patient records dating from the 1960's were used in this research, and the results were provocative. Nearly half the patients treated lived 2 years without a recurrence of the disease, compared to only ***20*** percent of patients in ***1960.*** Unfortunately, the change in therapy from 1960 to ***1970*** was also accompanied by other changes in diagnosis, treatment, and patients. The use of the computed angiographic tomography (CAT) scanner in the 1970's provided a much more sensitive test for detecting patients who did not have metastasis. At the same time, surgeons began removing metastasis in the lungs. At the Mayo Clinic, where both of these techniques were employed without chemotherapy, the survival rates equaled those of pa-

[3] R.I. Horwitz, and A. R. Feinstein, "Methodologic Standards and Contradictory Results in Case-Control Research, " *Amer. J.* Med. 66:556, 1979.

tients treated with the drugs. In addition, the patient mix probably changed over time so that those with the worst prognosis no longer constituted the majority of those treated. These criticisms of the research design and recent findings of a small controlled trial have convinced the National Cancer Institute to support a multicenter RCT to assess the efficacy of adjuvant chemotherapy for osteogenic sarcoma.

This design demonstrates the importance of *history* as a plausible rival hypothesis in interpreting research results. It also points out that innovations in medical technology are not discrete events, but are often accompanied by other changes in the organization and delivery of medical practice that can affect construct validity. For example, surgeons note that there were major changes in the procedure for CABG surgery in the mid-1970's (e.g., cold-blood technique) and attribute to these changes responsibility for the decline in operative mortality. But, as this discussion has shown, the decline could also be due to a corresponding change in patient mix as more low-risk patients were convinced of the benefits of this innovation.

Wortman, Reichardt, and St. Pierre (402) have also suggested multiple measurements as a method of strengthening the basic NECGD. They recommend "double pretests" to estimate the change in baseline behavior of subjects in the absence of any treatment (by allowing each person to serve as his or her own "control"). The double pretest considerably strengthens the basic NECGD and should be employed whenever there is time to conduct two pretests prior to treatment. It is feasible when there is time to conduct two pretests prior to treatment. It is feasible when there is some lag between patient application and acceptance in a treatment program, as sometimes occurs in oversubscribed programs with long waiting lists. In situations where treatment is or must be made immediately available, the use of a double pretest would probably not be consistent with professional ethics.

Time-Series Design.—Often, data relevant to the assessment of a medical technology are collected at regular intervals over an extended period. Data archives such as the one used in the Manitoba evaluation of tonsillectomy can provide periodic information on the frequency and outcome of an innovation. If this is the case, a time-series design can be used. This design consists of multiple observations prior to and subsequent to the initiation of a treatment or other type of intervention (see table C-3 top row). Analysis of a time series involves checking for changes in either the level or slope of the series after the intervention.

Using this design to study the impact of a hospital merger on a number of cost indicators, Whittaker (39I) was able to demonstrate that, contrary to prior belief, the cost-per-stay increased after the merger as did total expenses per patient day. Employing this quasi-experimental design and the sophisticated statistical analysis procedures that have recently been developed for it, Whittaker demonstrated that a complex "organizational" innovation had uniformly "unfavorable" impacts.

The interrupted time-series design may at first seem to be an attractive assessment methodology that coincides with a number of convincing innovations—e.g., renal dialysis and the cardiac pacemaker represent successful medical technologies that appear to fit this design. However, upon reflection, it is clear that other information was available and used in the assessment of these innovations—i.e., physicians knew what happened to patients who did not receive the innovation—they invariably died. In such cases where the prognosis or time course of a disease is well documented, the technology evaluator has the benefit of a comparison series: a multiple time-series design (see table C-3). The comparison series helps to eliminate a number of threats to validity (e.g., history and maturation) and to reduce the plausibility of others. Time-series data can provide useful and inexpensive monitoring of an innovation and can even furnish evidence of causal effects. Thus, they could be used in the postmarketing surveillance of medical innovations.

Statistical analysis of time-series data is still a rarity in the assessment of medical technology. Although

Table C-2.—Relationship of Methods and Policy Issues to the Innovation Process

Level of development	Method/validity	Policy issue
New	Needs assessment Technical feasibility/ construct validity	Social need
Emerging	Research design/ internal validity Cost-benefit analysis Secondary analysis/ statistical conclusion validity	Efficacy, safety, social impact
Existing	Postmarketing surveillance Data synthesis external validity Needs "reassessment" Cost-effectiveness analysis	Effectiveness, safety

SOURCE: National Institutes of Health, "Criteria for Identification of Candidate Technologies for Consensus Development," memo, Feb. 23, 1978

Table **C-3.—Multiple Time. Series Design**

0	0	0	0	0	X	0	0	0	0	0
0	0	0	0	0		0	0	0	0	0

SOURCE: Adapted from T. D. Cook, and D. T. Campbell, *Quasi-Experimentation: Design and Analysis of Research in Field Settings*, 1979.

there have been a few exceptions such as Albritton's[4] study of the 1966 Federal program for measles immunization, most researchers have been content to present their data in graphic form (see ***342).*** The effects of interventions are often dramatic, and visual judgments of statistical and medical significance may be adequate in many cases, However, statisticians (see **330)** have long warned that graphic representations of data can often be misleading. This warning has been specifically repeated with respect to time-series analysis (175,202). These authors demonstrate that visual and statistical analysis of time-series data often lead to opposite conclusions. More detailed descriptions of interrupted time-series analysis and examples of applications may be found in Glass, Willson, and Gottman (167), and McCleary and Hay (237). Recent tutorial articles such as the evaluation of a Regionalized Perinatal Care program in North Carolina (161) provide examples of the growing use of this design in assessing medical interventions.

Conclusions.—Although the cohort or NECGD is often easier to use than an RCT, it suffers several weaknesses in the form of threats to validity. The most serious threat are selection differences. Because subjects are not randomly assigned to treatment and comparison conditions, pretest or baseline differences among the groups are quite likely. These initial differences are then confounded with changes due to treatment observed at the posttest. A number of analytic approaches have been suggested to deal with this problem. For example, the Cox regression technique has been used to analyze the survival data from cohort studies (see 181). However, the analysis rests on the assumption of proportional hazards, that both groups have the same risk of illness, and this is unlikely to be true where the groups are nonequivalent. The results from such nonrandomized experiments thus remain extremely equivocal, particularly when the experimental and comparison subjects differ significantly in terms of important pretreatment characteristics.

Because there is no agreed upon analytical solution to the problem of baseline selection differences, probably the best that researchers can do currently is to use several different methods of analysis **(70).** If the results from the various methods are congruent, evaluators may state their conclusions with appropriate caution. If different methods lead to different results, the situation is more confusing, and the technology assessment will have to be more tentative in its conclusions. Although this design may be appealing, it poses such severe problems in analysis (i.e., it has doubtful statistical conclusion validity) that extreme care is warranted (402).

[4]R. B. Albritton, "Cost-Benefits of Measles Eradication: Effects of a Federal Intervention, " Policy *Analysis* 4:1, 1978.

In sum, do nonequivalent controls, particularly with matched groups, provide useful information for a technology assessment? Our general answer is that they do not. The methodological problems resulting from these designs are often of such serious concern as to undermine the credibility of the findings. Only when the competing explanations or rival hypotheses (i.e., important threats to validity) can be demonstrated to be implausible or can be ruled out through other subsidiary data should such studies be considered seriously in a technology assessment. Statistical solutions, in particular, should be viewed with skepticism despite the impressive impenetrability of their algebra.

As this discussion has shown, there is justifiable concern about the credibility of the evidence produced by controlled nonrandomized studies. For many experts and informed practitioners, the potential existence of such methodological problems is sufficient to cast doubt on the findings. These concerns in determining the efficacy of medical innovations are not new. Meier's ***(243)*** discussion of the Salk polio vaccine trial of 1954 indicates that the original quasi-experimental design was upgraded to an RCT in certain States because of these concerns. The original design called for second-grade children to receive the vaccine with first- and third-graders as comparison groups. The difference in the size of the effect observed from these two designs illustrates the problem with estimates of efficacy obtained from cohort quasi-experiments. The differences in the incidence of polio cases was ***40*** (per 100,000) for the RCT, while it was only 27 for the alternating grade cohort quasi-experiment. The nonrandomized results thus underestimated the efficacy of the vaccine by nearly ***50*** percent.

The multiple time-series quasi-experiment is much stronger on internal validity than the cohort design. The research by Sherman (342) and his associates (343) demonstrates the potential utility of the time-series design in the evaluation of health programs at the individual patient or program level of analysis. The time-series design is relatively unobtrusive; it rarely requires the changes in operating policy that a randomized experiment often does. In addition, time-series analyses may be used to assess innovations that have already been in operation for a considerable length of time (see 403). Since many observations are required to perform time-series analysis, it is most appropriate for agencies that collect data at regular intervals. Hospital and insurance reimbursement or claims records would be most useful, These could be used to provide information on cost, utilization, and health outcome.

UNCONTROLLED DESIGNS

The most common form of evaluative study for medical technology assessments employs a nonex-

perimental or uncontrolled design (401). These so-called case studies do not include any comparison groups at all and usually report judgments by physicians about the extent to which each patient improved.

The first studies on gastric freezing (143) were almost all of this type. The results of the early studies were largely on the self-reports of a few patients subjected to the procedure. The basic case study may be described as a "posttest only" design, since only one measurement of the status of the subjects is used. A slight elaboration of the "posttest only" case study is the one-group pretest-posttest design, which includes a measure of the status of the patients prior to, as well as after, treatment. The use of two assessments allows the researcher to estimate changes in the patients over the course of treatment, as well as their final status.

The problem in using nonexperimental evidence to assess a medical innovation is illustrated by a new technology to facilitate the management of diabetes—home blood glucose monitoring. This is a fairly recent innovation that shows promise of helping diabetics monitor their blood glucose levels more accurately than before, thereby allowing them to participate in their treatment by changes in diet and exercise ***(362).*** To obtain information on blood glucose level, the patient pricks a finger and applies the blood to a reagent or chemstrip. Glucose level can then be determined either directly or by reading a reflectance meter. This technique is viewed as a replacement for urine testing, although it costs three to four times as much.

There is a great deal of enthusiasm about blood glucose monitoring, and it is being introduced in a number of diabetes outpatient clinics. However, there is little evidence as to its effectiveness. Only nine studies of this innovation could be found, and none of them used a control group. Furthermore, many of the studies simultaneously introduced other regimens with the blood glucose monitoring procedure, thereby raising construct validity questions. The other regimens introduced included exercise, group therapy, and spray injection of insulin. Any of these techniques could have produced the beneficial effects reported. Moreover, most of the studies had very few patients; five had ***17*** or fewer subjects. Thus, selection of highly motivated patients, for example, could produce overly optimistic results.

Although nonexperimental studies can provide information concerning the technical feasibility of a new medical technology, they are far from definitive. They can also provide useful "qualitative" information ***(286)*** concerning the acceptability of the technology to patients (e.g., their willingness to draw repeated blood samples from their finger), factors affecting compliance (e.g., the interpretability of the chemstrip), and related behavioral issues that may hamper its utility. However, these studies should not be viewed as providing adequate information concerning efficacy and safety. Without a valid comparison group, it is not possible to determine whether the benefits are due to patient self-selection or to other factors. Nor is it possible to tell whether the innovation is superior to the urine testing methods now commonly used.

The major difficulty with nonexperimental designs is that they are subject to practically all of the threats to internal validity described above. It is inappropriate to interpret such studies as indicating that observed changes in patients are due to the innovation. Unfortunately, such interpretation is a common occurrence. Physicians, lacking training in research methods, can mistakenly perceive such preliminary pilot studies as being definitive. This can result in premature diffusion. The situation is often exacerbated by the exaggerated claims made by the developers of the technology.

Research Synthesis: Integrating the Results of Multiple Studies

The preceding section emphasized a set of principles underlying the design and interpretation of individual studies to assess the effects of medical technologies. The assessment of medical technology, however, is a process that involves more than the consideration of a specific research study (see ***266,269).*** In order to conduct a technology assessment, multiple sets of evidence, where available and relevant, must be considered and synthesized. Although little attention has been given to methods for synthesizing evidence about the effects of technology ***(266,398),*** there do now exist formal techniques to integrate the findings from different studies and to develop generalizations based on their results. The methodological and conceptual issues involved in the conduct of such analyses are considered in the discussion below as part of the technology assessment process.

The synthesis of research data is often both controversial and complex. Controversy arises because the results of studies about a particular technology may vary and/or be interpreted differently by different assessors. Synthesis is complex because medical technologies may have different clinical outcomes depending on who uses them or when they are used. Establishing the efficacy and safety of a technology on the basis of research evidence is typically a lengthy process. These assessments (i. e., safety, etc.) ***depend*** basically ***on*** the amount and quality of the research evidence and the analyst's ability to deal with the available information (i. e., ability to determine the

validity of the evidence and to combine the various types of information appropriately).

This section focuses on problems of synthesizing independent research studies relevant to a technology assessment and describes some formal procedures that enable systematic integration of research results. In addition to describing quantitative methods, it considers a number of methods for synthesizing information that rely on group decisionmaking approaches to technology assessment. These methods are often used to resolve controversies about research evidence and to develop guidelines for employing particular medical technologies.

Although information from a variety of sources must be considered as part of a technology assessment (e.g., costs, social impact, etc.), research data concerning efficacy and safety form the central component. These outcomes are essential for determining social impact (see ***295).*** Methods for synthesizing research information and using it in decisionmaking are critical to the outcome of an assessment. One purpose of the following discussion is to suggest what types of data are useful in the development of technology assessments and how they should be treated. The section is organized in four parts: 1) the application of the validity concepts introduced in the previous section to the synthesis of multiple research studies; 2) current approaches and problems to synthesizing research evidence; 3) quantitative research synthesis and integration methods; and 4) formal, group decisionmaking methods for synthesizing research evidence.

Validity

The previous section of this appendix emphasized inference problems inherent in the interpretation of individual studies of medical technology. From a methodological perspective, true experiments, particularly RCTs, reduce problems of equivocality of inference, as compared to other research strategies. A single RCT, however, cannot resolve all questions about a technology, and technology assessments cannot rely solely on their availability. If randomized studies are not available, decisions will have to be made about how to treat the validity problems inherent in other types of research. For example, the reduction or elimination of threats to internal validity by an RCT does not automatically avoid problems due to low external, construct, or statistical conclusion validity. In particular, external validity often must be established by examining evidence from multiple studies. Thus, validity considerations are as relevant to the problems of aggregating and synthesizing the results of many studies as they are to interpreting a single study. In the following discussion, the validity framework is extended to indicate its use in integrating evidence from multiple studies.

INTERNAL VALIDITY

Internal validity problems are central to the synthesis of findings from multiple studies. Because of the limited availability of RCTs, other evidence that can reduce the number of plausible alternative explanations for findings should be considered. However, the validity of nonrandomized studies must be carefully examined. If all, or most of the evidence about a particular technology was generated through similar, and perhaps consistently flawed research designs, the advantage of multiple sets of data may be lost. If the available literature includes a large number of studies with low internal validity, then a simple aggregation of the results may yield a conclusion open to a variety of alternative interpretations, especially when similar validity problems affect each study.

The existence of a few studies using randomized control group designs, on the other hand, does not guarantee high internal validity. Again, the alternative explanations must be considered to determine if they can be eliminated. For example, the previous section noted that the RCTs assessing the efficacy of CABG surgery were consistently flawed by differential patient attrition (or experimental mortality). Only through the availability of other evidence provided by additional control groups or improved analyses can the remaining threats to internal validity be eliminated (see 71). The principal problem in data aggregation is to identify such validity problems and to develop a strategy for aggregating the results of studies that differ in their internal validity.

In most cases, it is likely that experimental, quasi-experimental, and nonexperimental data will be available. The problem, then, is the appropriate choice of both the evidence and the amount of emphasis it should be given. One possibility ***(166)*** is to aggregate studies that are high in internal validity separately from more "poorly controlled" ones.

Gilbert, McPeek, and Mosteller ***(159,160)*** provide evidence of the importance of this strategy for medical technology assessment. These investigators compared the results of randomized and nonrandomized clinical trials of a series of medical innovations. They found that positive results were more likely to be obtained by an uncontrolled research study than by an RCT. RCTs tended to yield much less favorable conclusions about effectiveness. For example, among 53 studies of portacaval shunts, they found only 6 well-controlled trials. Of these controlled trials, three were associated

with negative conclusions about the treatment and three yielded moderately positive conclusions. This compares with **32** uncontrolled studies, where 24 were very positive, 7 were moderately positive, and 1 was negative. In general, Gilbert and colleagues found that the poorer the methodological quality (i.e., the lower internal validity), the more likely that a treatment would appear to be effective. The implication is that conflicting claims surrounding medical innovations may merely reflect differences in the validity of the research designs.

STATISTICAL CONCLUSION VALIDITY

In evaluating a set of studies, it is necessary to consider whether serious threats to statistical conclusion validity exist in individual studies and whether these threats prevent developing conclusions about the technology under study. Berk and Chalmers ***(26)*** examined the adequacy of the statistical analyses in a research synthesis of studies dealing with the cost effectiveness of ambulatory care (see below). They reviewed those studies reporting no difference in clinical outcome to determine whether there was sufficient statistical power to detect a 25-percent difference (if one existed). Of the ***23*** randomized trials, 16 had sufficient power. Seven RCTs plus all the nonrandomized controlled trials were classified as having "indeterminant clinical outcomes since selection bias may influence the outcome and obviate statistically valid comparisons when controls are not selected at random. "

EXTERNAL VALIDITY

External validity is essential in assessing data from multiple studies. The more widely a treatment has been tested, the easier it should be to establish the degree to which results are generalizable to various populations and settings. Studies high on internal validity, such as RCTs, may often yield differing and apparently conflicting results, because different patients, settings, or procedures are used. It is crucial that these studies be aggregated or stratified according to external validity factors. The differences can often be dramatic. The NIH consensus conference on CABG surgery ***(96)*** found the surgery effective for patients with left-main coronary artery disease, but not for patients with single- or double-vessel disease. Similarly, radical mastectomy, once the universally recommended procedure for breast cancer, is no longer endorsed by experts ***(96)*** for women whose disease is detected early (i.e., Stage I and II).

CONSTRUCT VALIDITY

Construct validity is also a serious concern in synthesizing the results from many studies. As Pillemer and Light ***(292)*** have noted, it may be that differences across studies are due to the use of treatments that have only been labeled similarly. For example, surgeons at the consensus conference on CABG surgery, noted above, dismissed most of the studies raising concerns about the procedure's safety (i.e., high operative mortality), because the studies were conducted before a major change, called the "cold-blood technique, " was adopted in the mid-1970's. The modification of a new technology can greatly affect its performance. These changes, which are often unreported, mean that an evaluation is, in fact, assessing a family of technologies, or a "moving target" (380). The rapid development that characterizes the early stages of technological innovation can lead to errors in data aggregation, because unreported, new components of the treatment may have been incorporated into various assessments. Often these labeling problems are more insidious in that other unnoticed technological changes co-occur with the innovation (e.g., improved diagnosis in osteogenic sarcoma).

Outcome Measures.-The major problem confronting those desiring more systematic methods for synthesizing the results from many research studies has been the inability to combine many different measures of efficacy and safety. One must ensure that comparable measures of the appropriate construct have been employed. For example, Berk and Chalmers (26) report a systematic review of the efficacy of ambulatory care as a cost containment measure to reduce inpatient expenditures. They found 134 relevant articles. Studies lacking either construct or statistical conclusion validity were eliminated. Of the **109** actual studies reported, 31 were eliminated because economic outcomes were not discussed. In the remaining ***78*** investigations, they found an appropriate measure of costs in only four studies! Thus, the improper measurement of a construct can significantly reduce the validity and usefulness of many studies for synthesis.

Problems With Traditional Synthesis Procedures

The traditional approach to synthesis is the literature review, Almost all technology assessments begin with such **a** research summary. Unfortunately, these reviews tend to be asystematic and subjective. Reviewers select the evidence they believe to be most relevant and typically organize their presentation around the dem-

onstration of a particular hypothesis. Although technology assessments, such as those developed by the former NCHCT, were based on summaries assembled and reviewed by several experts, there are still a number of problems in relying on this approach to derive the implications of research and to resolve controversies.

METHODOLOGY

A central problem in literature reviews is how to deal with methodological issues. As noted above, the relatively few RCTs generally available (401) present an important obstacle to the reviewer. Well-controlled research studies are probably the best way to produce unequivocal evidence, However, the weight of other evidence may sometimes hinder their use.

This problem is illustrated by the controversy over electronic fetal monitoring (EFM). Several recent reviews of the efficacy and cost effectiveness of EFM (e.g., 17,367) have indicated that significant risks are associated with monitoring (in particular, an increase in cesarean section rate) and that it is not a beneficial diagnostic tool for many of the patients with whom it is being used. Significantly, reviewers who are skeptical of the use of EFM are primarily researchers; reviewers who are clinicians have come to a different conclusion and have strongly supported the broad use of EFM (see, e.g., 189). Researchers appear to disregard much of the published literature because it consists of reports of uncontrolled research. Wortman ***(401)*** notes that 23 of the 24 poorly controlled studies supporting EFM--there were no well-controlled studies supporting it—employed historical controls. From the perspective of a research methodologist, the lack of internal validity indicates that there is no valid basis for comparing monitored to unmonitored births. Clinicians, in contrast, appear to be swayed by the large number of case studies that describe successful applications / assessments of EFM. Since the rate of false positives leading to cesarean section is relatively low, this literature probably is most consistent with their own experience.

Even when RCTs are available and the weight of the evidence is not as discrepant as in the EFM situation, they may not fully answer questions about the technology. Tonsillectomy is a case in point. A substantial literature exists about the safety and efficacy of tonsillectomies, and experimental, quasi-experimental, and nonexperimental research is available. Cochrane **(60)** reports three different clinical trials on tonsillectomies conducted in England during the 1960's, but he contends that none of the trials resolved the policy controversy over the appropriate use of tonsillectomy. According to Cochrane, the available RCTs exhibit two methodological problems: 1) the treatment was compared with no or inadequate medical treatment (instead of an alternate treatment); and 2) the patients' parents were not blind to the conditions of the experiment, so those whose children were on the waiting list may have exaggerated their children's symptoms.

Wennberg, Bunker, and Barnes ***(389)*** note that a large-scale clinical trial is currently being conducted, but that the trial, in itself, will not resolve the controversy. This is because the current RCT does not include a sample of the full population of children for whom tonsillectomy is recommended. In essence, several internal and external validity problems prevent these available and pending RCTs from being unequivocal tests.

TIMELINESS

Some (e.g., 34) believe that clinicians, over time, will be able to determine which medical treatments are useful and which are not. The implication is that methodological considerations are not central. Others (e.g., 389) have suggested that this approach is ineffective and that many common medical practices are inadequately evaluated and perhaps worthless or unsafe. A question exists as to whether systematic reviews of research evidence can influence medical practice.

Several studies have examined the use of research by clinicians. Fineberg, Gabel, and Sosman (145), for example, reviewed the use of scientific papers by anesthesiologists. They found that there is a significant lag between research discoveries and their publication. Their view is that scientific papers affect actual practice slowly. This would seem especially true of reviews that attempt not only to summarize but also to draw implications from the literature. In part, this is because it takes considerable time for a published literature on any medical technology to develop. In several now-classic cases (e.g., gastric freezing), literature reviews were only published years after a procedure was abandoned because it was ineffective or unsafe (see ***143, 245).***

In the gastric freezing example noted above, either a more timely RCT (i.e., earlier) and / or more systematic attention to the available nonexperimental data might have hastened the abandonment of the procedure. These two evaluative processes are, in fact, related. Thus, if an RCT is not conducted during the initial investigational stage of a developing technology, then it is even more important that systematic attention be given to whatever data are generated, since these data may indicate whether or not an RCT is needed. It should also be clear that an RCT may not "solve" the technology problem, and, in many cases,

research synthesis may stimulate the generation of additional data that are needed.

CONCLUSIONS

The evaluation of safety and efficacy evidence to understand the effects of particular medical technologies is complex. Complexity is related to the presence of bias and methodological problems, such as the lack of appropriate control groups in research reports and literature reviews. Research evidence may exist for any medical technology, but it may be difficult or impossible to synthesize these data without carefully considering the validity of the individual studies. Developing research that can be used for a technology assessment is obviously difficult, but is only the first step. Synthesis strategies are clearly necessary as part of this process to deal effectively with the results of the many studies bearing on a technology.

Systematic Procedures for Data Synthesis

A major implication of the previous discussion is that the need for policy-relevant information often outstrips the capabilities to provide it. The development of conclusive evidence about a technology at present seems to be a relatively slow process. This discovery process is probably better geared to the occurrence of "breakthroughs," those rare single studies or programs of research that resolve a controversy, than to dealing with elaborate arrays of potentially conflicting or inconsistent information. Procedures are needed that enable the accumulated insight gained from research to be usable within the technology assessment process.

The problems and benefits in systematically organizing and integrating research findings are discussed below. The procedures described, although not representing a panacea for all the problems identified, suggest how the process of research synthesis can be more rigorous. Some elementary qualitative procedures, as well as sophisticated statistical techniques, for conducting research synthesis are described below. The goal is to outline the range of systematic methods that may be employed and to contrast them with more traditional techniques.

VOTING METHOD

A simple form of synthesis has been called the voting method (226). This technique essentially involves organizing a body of literature according to some prespecified set of criteria. Usually, vote counting involves the selection of a particular sample of outcome studies, coding some aspects of their design and/or conceptual framework, and classifying the observed outcome(s) according to whether they are favorable, neutral, or unfavorable (i.e., "taking a vote"). The Gilbert, McPeek, and Mosteller (160) study, referred to above, is an example of this type of synthesis. Sampling the literature to determine the rate of successful innovation in anesthesia and surgery, their analysis indicated that about half the innovations assessed by RCTs were successful when compared to a "standard" treatment.

A frequent use of the voting method is to demonstrate differences obtained by various methodological approaches. For instance, Gifford and Feinstein (157) critiqued studies of anticoagulant therapy for acute myocardial infarction (MI). They examined all available literature on acute MI that reported control group studies of acute MI treatments. For each of **32** studies located, they coded the degree to which the diagnostic criteria for MI were clear, whether randomized control groups or other methodological criteria were employed, and summarized their findings in several contingency tables. The results of the vote count indicated that anticoagulant therapy was superior to no treatment more often in reports that did not observe methodological standards than in those that did.

The strength of vote-counting analyses lies in: 1) the precise identification of the populations of studies to be sampled, and 2) the coding of substantive and methodological aspects of the study according to clearly defined procedures. More widespread use of the technique could probably aid in determining which specific patient populations and/or conditions could be effectively treated by a medical technology. The voting method helps to avoid the problems of reviews that only selectively describe research or pay attention only to some aspects of the study. In addition, such analyses may be particularly useful in identifying relationships between methods and outcomes.

Krol (216) cites three problems with the voting method: sample size, effect size, and Simpson's paradox. Large studies are likely to produce statistically more significant results than those with small numbers of subjects due to differences in statistical power. Thus, a finding of no difference among treatment and control conditions will be correlated with small sample size. In fact, Hedges and Olkin (186) have demonstrated that the voting method itself generally lacks statistical power. A second problem is the all-or-none nature of the method, Some findings may show small, marginal effects and others large ones, but they would count the same. Consider the case where effect size is correlated with outcome—large, positive *effects and* small, negative ones. The voting method would yield no difference when, in fact, there was an overall positive effect. Simpson's paradox is a more subtle statis-

tical point in which it is possible, under certain conditions, to reach different conclusions by aggregating data from each study rather than by counting each study separately. The paradox results from unbalanced cell frequencies. Finally, Light and Smith (226) have noted these and some additional problems with the method. The most important is that vote counting may oversimplify the results of studies and cause one to overlook more subtle, but important, relationships (especially interactions among variables).

META-ANALYSIS

A second synthesis technique, called "meta-analysis," has been developed by Glass (164,165). Meta-analysis or the "analysis of analyses" is a rigorous statistical approach to research synthesis. Meta-analysis utilizes the actual results of studies and permits the determination, across a set of studies, of the magnitude-of-treatment impact. Most statistical analyses, as summarized in research reports, ignore both the size and direction of effects and yield only a global probability of a "significant" difference. Meta-analyses are useful for assessing treatments where a large number of studies exist and where findings across studies seem to have great variability. As used by Glass, such analyses require that comparison groups be available (i.e., either randomized or quasi-experimental groups) and that the original research reports contain appropriate statistical information such as the group means and standard deviations. Glass (164) describes some indirect procedures for deriving the effect size from the inferential statistics reported in a study (i.e., t-test, F, etc.).

Effect sizes (ES) are calculated by determining the difference between the mean of the treatment group (T) and the mean of the comparison group (C), divided by the standard deviation of the comparison group (SD_c). Thus,

$$\frac{\overline{T} - \overline{C}}{SD_c}$$

This procedure converts the average effect of each outcome measure into a common scale (i.e., standard deviations) that can be compared to results of other studies. If a treatment has no effect, then there would be a zero effect size; if the treatment is effective (i.e., better than the current alternative), the effect size is positive; and, if the treatment is inefficacious, the effect size is negative. By making some assumptions about the skewness of experimental and control group scores within each study, and the distribution of effect sizes across a large number of studies (i.e., that they are normally distributed), effect sizes can be converted into percentile ranks and inferences can be made about the overall effects of a medical technology.

One of the best recent health technology examples of a meta-analysis is Smith, Glass, and Miller's review **(354)** of the outcome studies of psychotherapy treatments (see also, 353). Smith and colleagues searched the published literature, including abstracts, and included within their analysis all available control group studies of the effectiveness of any form of psychotherapy. Drug studies were analyzed separately, while those studies that did not involve the use of professional therapists (operationally defined as psychologists, psychiatrists, and social workers) were eliminated from the analysis. The investigators coded an extensive number of variables for each study, including methodological criteria such as the nature of the patient assignment to condition (e.g., random v. matching), experimental mortality, and other threats to internal validity. Effect size scores were calculated for each principal dependent measure. The analysts also developed a code for validity of the outcome measures.

Smith, Glass, and Miller's ***(354)*** findings indicated that, on the average, the difference between scores of the groups receiving psychotherapy and scores of the control groups was ***0.85*** standard deviation units. Assuming the normal distribution of effect size scores, this average standard score indicates that a typical person who receives psychotherapy is better off than **80** percent of the people who do not. Smith and colleagues also conducted a number of analyses to determine whether the methodology of the study affected results and whether different therapies (or other factors) were differentially efficacious. They found few reliable methodological differences. It appeared that outcomes were not related to the use of randomized control groups. This finding should, however, be tempered by the knowledge that all of their sample studies used comparison groups and were generally high in internal validity. When this is not the case (i. e., where quasi-experiments are included), then the outcome can vary with the methodology (i. e., research design). Wortman and Yeaton[5] have shown this to be the case for the studies on CABG.

There has been some criticism of Smith and Glass' (353) approach based on their "lumping together" of a large number of what some consider incomparable treatments and outcomes (e.g., see 137). The strength of the effect size technique, however, is that it provides a common metric that permits analysis of the differences (methodological and substantive). Smith and colleagues' classification variables for each study were fairly comprehensive and yielded a systematic comparison of studies on the basis of their conceptual and methodological designs. What is problematic

[5] P. M. Wortman, and W. H. Yeaton, "Synthesis of Results in Controlled Trials of Coronary Artery Bypass Graft Surgery" (Ann Arbor, Mich: Institute for Social Research, 1982) (report submitted for publication).

about such meta-analysis, however, is that the findings are heavily dependent on a number of decisions that are not always made explicit. These include the studies selected/rejected from the literature, variables included/excluded, and their construct validity. It is not possible to ascertain biases resulting from Smith and colleagues' sampling decision nor whether only certain types of studies, therapies, or variables are assessed using control group designs ***(273).*** A broader analysis of psychotherapy research might yield different conclusions than those drawn by these investigators.

OTHER SYNTHESIS TECHNIQUES

A number of other methods exist for statistically combining the results of independent studies (see ***69,292,324).*** The effect size method described above actually incorporates several procedures. The most important of these methods is the comparison of treatments to detect interactions between characteristics of a study and outcome (i. e., external validity issues). As noted in the earlier discussion of the voting method, some of these procedures can be employed when effect scores are not computed. Additional statistical methods combine probability values from various studies and adjust outcome scores according to the relevance of the data.

Rosenthal (324) describes a number of procedures for combining probabilities. These range from adding observed probability (p) levels across different studies to adding weighted standardized (z) scores. These methods also include the testing of mean probability values. Essentially, using such procedures allows one to indicate whether significant effects are obtained across a set of studies. The problem in using probability values is one of statistical conclusion validity. The number of subjects per study influences the statistical power to detect whether significant overall differences are present.

DuMouchel and Harris (131) discuss another interesting quantitative method for synthesizing the results of experiments done with human and animal species. This method, a sophisticated application of Bayes' theorem, provides estimates of carcinogenic risk from various substances derived from the results of epidemiological studies.

IMPORTANCE OF SYSTEMATIC DATA SYNTHESIS

Earlier, it was noted that technology assessment is essentiality a synthesis process that involves the review and integration of research findings. There are a number of specific benefits that result from employing formal procedures for data synthesis (see 292). The first advantage is that formal syntheses help to identify contradictions in the literature by systematically organizing studies according to specified classification factors. It becomes possible to segregate differential outcomes according to treatment characteristics and/or methodological approaches. The analysis of different findings when controlled and uncontrolled studies are employed (see 160,400) is a good example of this aspect of meta-analysis.

A second benefit of meta-analysis has to do with the use of effect size scores. Not only do such scores provide insight as to the worth of the treatment, as in the Smith, Glass, and Miller **(354)** psychotherapy example, but they also provide a benchmark for later research. Thus, for example, a meta-analysis conducted by **Posavac (294) of 23 controlled studies of patient education programs found a 0.75 average effect size. Posavac indicates that this should provide a standard against which new patient education programs can be assessed. If the effect sizes of new programs** are only 0.20 (and similar dependent measures are employed), this would probably indicate that the programs are not particularly effective, at least for the problem or population for whom they were designed.

Another advantage of quantitative synthesis methods are that they serve to control for certain statistical conclusion validity problems (e.g., power) that some commentators have reported as severe in the medical literature (e.g., 141,172,337). It can be assumed that the widespread use of meta-analysis and other quantitative approaches to synthesis would improve statistical reporting practices by calling attention to different investigators' use of data. In addition, errors in analyses, such as the use of multiple independent inferential tests without appropriate error rate control or incorrect inferences because of a lack of power, would be compensated for by most meta-analytic procedures. Although errors in data collection and, perhaps, in computation of means and standard deviations would not be corrected by these synthesis methods, the systematic analysis of multiple studies should render the effect of such errors less consequential. The attention to systematic considerations of the "weight" of evidence across research studies should have a general salutary effect.

Finally, it should be noted that, although these procedures seem most appropriate for evaluating more mature technologies that have accumulated a considerable body of research, they are often applicable to less developed technologies. In some cases, where only meager evidence is available from a small set of studies, it may be that a review of specific components from some other portion of the literature may suggest the effectiveness of the new technology. Thus, physiological evidence may be considered with other clinical,

experimental data as in the case of radical mastectomy (see 147) noted earlier.

Group Decision Methods

Although the application of formal statistical procedures for the integration of data from individual studies should improve the ability to conduct technology assessments, the use of such methods does not entirely resolve policy controversies. Such analyses cannot go beyond the available data on a particular problem, nor can they substitute for informed judgment. In the discussion below, some recently suggested procedures for resolving conflicts across research studies and for developing assessments of particular technologies are described. These informal methods include a new approach to decisionmaking sponsored by NIH, referred to as consensus development, and a number of other decisionmaking techniques (e.g., Delphi) that have been employed in assessments of medical technology.

NIH CONSENSUS DEVELOPMENT

In response to congressional pressure to assist in the transfer of technology, NIH initiated its consensus development program in 1977 ***(310).*** Perry and Kalberer **(287)** recently described the consensus development program at NIH. Its goal is to bring together various concerned parties (e.g., physicians, consumers, bioethicists) in order to seek agreement or "consensus" on the safety, efficacy, and appropriate conditions for use of various medical procedures. Judgments about the technology under consideration are intended to be based on the scientific evidence of its effectiveness as well as on information about its social, ethical, economic, and legal impacts. The consensus development process is designed to produce a written recommendation, called a "consensus statement, " that can be accepted by clinicians and researchers. The statement is supposed to identify both what is known and not known about the technology.

Topics for NIH consensus development are chosen because of their current or potential importance (e.g., in terms of cost, number of patients affected). Since September **1977,** NIH has held more than **30** consensus conferences at which the evidence and implications of a wide variety of technologies have been considered. Topics have ranged from bee sting kits to CABG surgery. The technologies include both emerging, as well as currently used, technologies that either have not been carefully evaluated for safety and efficacy or are controversial. Recently, there has been a trend toward more mature technologies (see next major section below) for which there is more scientific evidence concerning effectiveness.

Over the past few years, the conferences have generally followed a similar format. A panel of neutral experts is selected by NIH to hear presentations by the leading medical researchers addressing a prespecified set of questions about the technology. The presentations, usually summarizing the latest research findings, are made over a 2-day period during which both panelists and audience members discuss the research findings. On the evening of the second day, the panel is sequestered to draft a statement responding to the questions. Usually, they deliberate through the night, writing as many as four drafts of the consensus statement. In some rare cases, minority reports are developed to indicate disagreement with the majority recommendations. The next morning the statement is read to the audience for their comments and criticisms. The conference concludes with a press conference. The panel then disperses with the final task of revising the statement. The consensus statements are widely disseminated by NIH through direct mail to thousands of organizations and individuals and by publication in leading medical journals such as the *New England Journal of Medicine* and the ***Journal of the American Medical Association.***

From a methodological perspective, two aspects of the consensus development process are of concern: 1) its sensitivity to the limitations of the research evidence, and 2) the extent to which a comprehensive and systematic review of the research literature is considered. There is little published evidence concerning these issues. An examination of panelists participating in previous consensus conferences***(96,97,98)*** indicates that there has been no consistent policy to include a methodologist—either a biostatistician or epidemiologist. On few panels were such persons included. This means that in most cases there was no informed person who could indicate the methodological limitations of a study. The problems to which methodological ignorance can lead have already been described.

The consensus conference on CABG surgery was an exception (95). Two biostatisticians are listed as members of the panel, and their influence on the consensus statement is evident. The methodological limitations of the research literature with respect to a key question are discussed at length (see 95). A number of these methodological problems have been noted above: attrition due to crossovers, use of historical controls, statistical analyses of registry (i. e., quasi-experimental) studies, and the like.

Despite this indication of methodological detail, there is apparently no formal policy to provide syste-

matic reviews of the research literature. For example, the weight of the evidence on the efficacy of the coronary bypass procedure as presented in the published consensus statement (95) was evidently derived from two large, multicenter RCTs: the somewhat controversial VA study (255) and the ongoing European trial (136). Our examination of the literature revealed that there are at least 30 studies, of which 9 are RCTs. Given the emphasis on external validity issues (i.e., identifying the patients for whom the surgery is beneficial), the limitation of the discussion to two studies was clearly unwarranted.

This problem has occurred in other consensus conferences as well. In a recent letter to the Journal of the ***American Association,*** Jones ***(203)*** noted that one of the conclusions from the conference on adjuvant chemotherapy of breast cancer was "based on incomplete information. " He pointed out that the results of only five studies were presented. while there were "at least nine major studies" containing "convincing evidence" on the effectiveness of chemotherapy in postmenopausal women. (The consensus statement claimed effectiveness for only "a select group of breast cancer patients.")

On the other hand, in most consensus conferences, the attention to these methodological concerns has been reversed. Most consensus statements reveal little discussion of methodological issues and limitations of the studies even where this might be appropriate. However, extensive background materials are often made available to the panel. These included a computerized bibliography of the literature and reprints of the articles.

The consensus conferences are coordinated by NIH's Office for Medical Applications of Research (OMAR). Although the topics are selected by the relevant institutes, OMAR makes the final decision about the suitability of the topic, panel composition, and the proposed format for a consensus conference. Over the past 2 years under OMAR's direction, the conferences have developed in a number of ways. The use of a fixed format has already been noted. Other approaches involving adversary (i.e., nonneutral) panels and task forces have been almost entirely abandoned. Moreover, the questions that have been posed to the conferences have been addressed strictly to those issues on which there is enough factual evidence to reach agreement. This has resulted in the omission of controversial issues. For example, in the recently published statement from the Reye's Syndrome consensus conference (67) questions about the role of salicylates (i.e., aspirin) were deliberately omitted because OMAR felt little was known about it (although the limitations of the studies establishing this association were briefly discussed). An editorial on the coronary bypass consensus statement in the *New* ***England Journal of Medicine (308)*** complained that it and other consensus statements "represent the lowest common denominator of a debate—the only points on which the experts can wholeheartedly agree." This reflects the current orientation of OMAR away from "state-of-the-art" conferences. One methodological consequence is that gaps in knowledge and needs for further research may not be as readily identified.

FORMAL GROUP DECISION METHODS

In addition to the NIH consensus development process, a number of systematic procedures for developing consensus based on behavioral science principles (see, e.g., 163) have been developed. The goal of these procedures is to aid groups composed of individuals with different information and perspectives to develop group judgments that best take account of the positions of the individual members. In the discussion below, two methods—Delphi and nominal group technique (NGT)—are presented. These techniques illustrate the potential and limitations of these methods for technology assessment.

Delphi Technique.—Delphi (78) is probably the oldest structured model for involving groups in decisionmaking processes and has been used widely in health care. The Delphi technique uses a series of questionnaries (or individual interviews), each followed by anonymous feedback summarizing all the participants' responses. Although Delphi was originally developed by the Rand Corp. to synthesize expert opinions on national defense problems, it has been extended to medical problems (232,246,250,318,336).

A unique feature of the Delphi technique is that persons selected to participate in the process generally have no direct contact with one another. Instead, participants are provided with a summary of the questionnaire responses, usually by mail. Personality or status variables, thus, have little chance to exert influence on a member's opinion, as they might in face-to-face meetings such as the NIH consensus development conferences, By using anonymous feedback, each expert has an equal chance of influencing other participants (41). The technique is also viewed as providing a framework within which to approach the problem in a focused manner. Finally, and perhaps most importantly, the technique provides a limited time frame in which to achieve consensus (41,135). There are a fixed number of iterations, usually three, in the questionnaire feedback process.

Delphi has been used to estimate the probability of an epidemic occurring. Information about morbidity

and mortality rates for both the total population and a high-risk population were sought in an investigation reported by Schoenbaum, McNeil, and Kavet **(336).** The investigators employed a modified version of the Delphi technique using two separate groups of participants. The first group consisted of five experts on influenza epidemiology and virology. Subsequent questionnaires fed back anonymous responses of the participants to the previous questionnaire. The second group consisted of 10 experts in immunization, infectious diseases, and preventive medicine. Their subsequent questionnaires were accompanied by summaries of responses compiled from previous questionnaires. The iterative process was continued until median estimates for each group varied by less than 10 percent from the previous questionnaire's responses. Since results of the Delphi process indicated that the probability of a full-scale epidemic was minimal, subsequent economic analyses revealed that it would not be beneficial to attempt to vaccinate the total population. They concluded that efforts should be directed at immunizing the high-risk population.

The Delphi technique has been criticized as being little better than the "seat-of-the-pants" method currently employed by policymakers, and as being a method which bases "knowledge" on an informal set of opinions rather than on formal decision analysis (332). Others (10) maintain that it is as subject to the same total error found in most predictions. The process is also time and group dependent, since the results are based on the information available to a specific group of experts at a specific point in time. It should be repeated as data change with time. It also appears less well-suited than face-to-face group meetings as a process for resolving minimally controversial issues **(318)** or for synthesizing the state of the art in a given field (163). Nonetheless, the technique's relevance for gathering predictive information seems clear (77). The Delphi technique may also have use in resolving highly controversial issues likely to be distorted when participants interact personally with one another.

Nominal Group Technique.—In another structured group process, members engage in limited interaction. Typically, all participants may be seated at a common table and asked to write their views on each of a number of issues posed by the leader of the meeting. Each view is recorded on a separate card, and talking is prohibited. The cards are collected, and their contents are listed for all to see without any indication of who is the author of each. The group then discusses these items, often choosing the ones that interest them most. Delbecq, Van de Ven, and Gustafson **(82)** call this the "nominal group" technique (NGT) because the individuals at the table (at the outset) are a group in name only. The (silent) presence of others while writing the cards creates social facilitation which stimulates participants to do well. Subsequent discussion dwells on the ideas proposed without any likelihood of distraction by attitudes toward those who did the proposing,

Thornell (368) has recently reported a study comparing the Delphi technique with the NGT. Physicians were randomly assigned to one of three Delphi or NGT panels to develop procedures for handling four hypothetical emergency medical services cases. In order to determine the reliability of the decisions, panelists were contacted individually **6** months later and asked to cast an anonymous vote on the procedures originally discussed. The degree of consensus achieved was the same for both techniques. The most striking finding, however, concerned the reliability of decisions over time. There were "very extensive" changes in the NGT vote ***6*** months later, suggesting that it is "a less than reliable technique for reaching a consensus. " In conclusion, although the physicians reported that they liked the NGT much more than Delphi, group norms and pressures were developed with the NGT that produced unstable or false consensual agreement.

Relationship of Assessment Methods to Stages of Innovation

In considering various methodological approaches to medical technology assessment, there are two related issues that must be examined, The first is how to deal with the limited funds available for conducting technology assessments. The second is when or where to intervene in the innovation process. In order to allocate scarce methodological resources, it is necessary to understand some essential properties of technological innovation in medicine.

There are many excellent examples of medical innovations in both the private and public sectors (e.g., 143,252,259). A very recent one—the portable insulin infusion pump—illustrates a number of generic issues in the innovation processes. The case study approach is limited (as was noted above); thus, this example is meant only to be descriptive rather than definitive.

Although the discovery of insulin as a "cure" for diabetes was a major breakthrough, subsequent experience with treatment by subcutaneous injection has revealed that it does not eliminate morbidity or mortality. Currently, diabetes ranks third among major diseases as a cause of death in the United States (309). Moreover, it is associated with a large number of crippling and debilitating conditions. For example, it is the leading cause of blindness. It also leads to myocardial infarcts, strokes, and other serious conditions.

Recently, there has been much discussion of and research on the possibility of using a portable insulin infusion pump to administer and control diabetes **(290).** Several investigators have demonstrated that such devices control not only blood glucose levels but other metabolizes as well (291). Although the exact cause(s) of the various pathologies associated with diabetes are still not understood, the results are nevertheless viewed as significant (133). However, these studies involve few patients—seven and eight, respectively—and should serve only as vivid case studies of the ability of these portable devices to achieve rapid and "strict control" over abnormalities associated with diabetes.

Although the underlying processes of diabetic-related diseases are unknown, it is believed that microvascular injuries (i.e., diabetic microangiopathy) result from the inadequate control obtained by conventional methods, primarily injection. There is now some provocative evidence (133,290) that the strict control obtained with the infusion pump can prevent and perhaps reverse these complications. Thus, there exists a physiological basis or "hypothesis" for the potential efficacy of this device. Such a physiological explanation is often essential for generating interest in a medical technology. When coupled with powerful demonstrations of potential efficacy such as were noted above, a technology possesses the essential ingredients for rapid diffusion.

Two other considerations also figure into the process. The first concerns the safety of the device; the second its availability. As noted above, diabetes is a major threat to human life and well-being. Bunker, Hinckley, and McDermott **(40)** observed in their review of a number of surgical innovations that under these conditions "efficacy is apt to be considered self--evident." It also appears that safety is seen as nearly negligible in such life-threatening situations. Where there is no alternative treatment and death is the likely outcome, patients and their physicians are motivated to try any promising innovation **(382).** Under such circumstances, innovations are likely to diffuse and diffuse rapidly. All that is required is sufficient availability or supply of the device. The literature on the infusion pump reveals that there are many manufacturers. It can thus be predicted that this technology is on the theshold of diffusion. Despite the many unanswered questions concerning the long-term effectiveness and acceptability of the pump, despite researchers' claims that it is "an experimental procedure which is still far from being a safe treatment routine, " and despite doubts about its effectiveness **(350),** the ingredients for the rapid diffusion of this technology are all in place.

Type of Technology

Throughout the preceding sections of this appendix, there has been an implicit assumption that the methods described are appropriate for all medical technologies. Is that assumption true? As shown in the preceding discussion, it applies to drugs and surgery, but what about devices, especially those involved in diagnosis?

OTA **(266)** has described five criteria for assessing diagnostic technologies, one of which is impact on "patient outcome. " This criterion has been the emphasis of the methods described in this appendix. Thus, diagnostic devices do not differ in their appropriateness for the methods for technology assessment discussed above. They only differ in the number of other criteria that can be used in their assessment (e.g., accuracy) and in the range of health outcomes they affect.

For example, two of the criteria OTA describes deal with the quality of the information the device provides. This involves established concepts and measures such as specificity and sensitivity of a diagnostic test (see 241). The other criteria deal with the organization and delivery of health services. These are important secondary impacts that should be considered for all technology assessments after the primary determination of efficacy and safety have been made. Banta and McNeil (15) provide an instructive example of these assessment criteria applied to the CAT scanner. They acknowledge that it is difficult to study health outcomes for this type of technology and also difficult to conduct randomized studies of it. As a consequence, secondary impacts involving nonrandomized studies using other criteria may be necessary in the short run. As previously noted in this appendix, such technology assessments require extreme care and cautious interpretation.

In addition to diagnostic, preventive, and therapeutic technologies, OTA **(269)** considers "organizational" innovations as a major category. Many innovations in health are primarily organizational in their medical function. For example, intensive care units (ICUs) represent a largely organizational change aimed at containing costs by centralizing patient care. Health planners and administrators, in particular, often regard ICUs primarily **as an** organizational change and not as a well-defined treatment with specified impacts. As Russell (331) notes, it has been "difficult to design a convincing test of intensive care's effectiveness. " The confusion between organizational change and health impact has also characterized the movement toward Professional Standards Review Organizations, health systems agencies, and many other major Federal health initiatives. There clearly is a need for planned innova-

tion where the rationale underlying change and its intended impact(s) are specified. The research designs discussed in this appendix are also applicable to these organizational innovations. However, much more attention needs to be given to the implementation processes or operation, and to the integrity of the innovation **(340).** Thus, it would be important to determine whether emergency medical services have been properly installed before assessing their effectiveness.

A Preliminary Strategy for Assessment

Our discussion of innovation raises the question of its relationship to the various methodologies described in this appendix. A number of researchers, including Williamson (396) and McKinlay,[*] have described models or stages in the innovation process that can be used to relate issues in validity and design to the development of a medical technology. According to a recent NIH conception similar to Williamson's (see table C-2), a medical innovation goes through three stages of development. At the earliest level (i.e., "new"), there is the perception of need such as a cure for a disease or a better way of diagnosing it and a preliminary assessment of the technical feasibility of the idea underlying the innovation (i.e., construct validity). The technology then becomes a reality, usually in an early form (i.e., "emerging") that can be assessed for its efficacy, safety, and social impact (e.g., quality of life). At this point, research design and validity issues (i.e., internal and statistical conclusion validity) as well as cost considerations are important. Once satisfactory evidence is obtained at this level, the innovation develops to an "existing" level where the emphasis is on its acceptability or external validity. Widespread diffusion of the innovation should occur at this point, and the relevant policy issues concern the cost effectiveness of the technology and the continued observation or postmarketing surveillance of the technology for unintended negative side-effects (388). Given the inability to predict the future impact of technologies—especially low-frequency, unanticipated negative side-effects such as toxic shock syndrome—continued surveillance using epidemiological (i.e., case-control) and related methods will be necessary.

The large number of potential technologies to assess and the pressures to develop and diffuse them quickly ensure that some stages of development will not be scrutinized with the appropriate methods. Many of the above examples (e.g., CABG surgery, gastric freezing) illustrate this point. In fact, it is the overdiffusion of young technologies and their associated costs that have led to the need for strategies to deal with the problems of technology assessment. The model described above, coupled the methods presented, provides the basis for such a strategy. There remains a need to order the technologies according to their priority for systematic, thorough assessment.

According to the model, technologies in the first stage of development do not need to be assessed. Since many, if not most, medical innovations will not progress beyond this point, the burden of assessment will be considerably reduced. Technologies maturing beyond this level can be ordered by the potential benefits and harm they pose. This ordering could be determined simply by calculating the product of the benefit or risk the technology poses to either decreased or increased mortality multiplied by the amount of use envisioned for the technology. For example, CABG surgery may pose a 4 percent risk of death for the 100,000 patients operated on last year. This would result in 4,000 deaths. Another decision rule could involve cost. Obviously, medical technologies could be ordered by both of these rules. The choice among the various possible ordering procedures is one that falls in the policy domain and is beyond the scope of this discussion.

Conclusions

A brief examination of innovation in medical technology reveals that it is a dynamic, temporal process that requires considerable flexibility in the methodology used. Different approaches are relevant at different stages of technological development. Moreover, policy-relevant evidence may not be available when needed, either because of the pressures for diffusion or the low priority for assessment initially assigned to the innovation. The Dalkon Shield, an intrauterine device, is a recent, unfortunate example of premature diffusion. Furthermore, no matter how thorough the assessment of a medical technology, there is always the possibility that unanticipated negative side-effects will be discovered at a later date when use of the technology is more widespread (e.g., X-ray treatment for facial acne). At such times, a decision to reexamine the technology will have to be considered along with the choice of an appropriate methodology for accomplishing this. Such postdiffusion technology assessments are much more difficult to accomplish. The recently initiated RCT to assess chemotherapy as a treatment for osteogenic carcinoma is an example of this process of surveillance and reassessment.

Given the scarcity of resources, it is unlikely that there will be much increase in the number of large-scale RCTs. Most of these require Federal support, and considerable funds are already allocated for such technol-

[*] J.B. McKinlay, "From 'Promising Report' to 'Standard Procedure': Seven Stages in the Career of a Medical Innovation," *Milbank Mere. Fund Q.* 59:374, *1981.*

ogy assessments. For example, Levy and Sondik (222) report that in 1976 about one-eighth of the total budget for NIH's National Heart, Lung, and Blood Institute—over $50 million—was devoted to major clinical trials. However, for many innovations, small-scale, local RCTs are probably feasible. Unfortunately, these are not often conducted. One can only speculate as to the reasons for this. Physicians often lack the methodological training to conduct such studies or the conviction that single-site studies are useful. The implication for medical technology assessment is that there will be an increased reliance on studies using other methods of evaluation unless some new policy initiative (see **307)** is taken. As noted in this appendix, these other evaluative designs are most vulnerable to challenge and are often seriously flawed. Where such quasi-experimental approaches are employed, replication and "triangulation" (71)—the use of multiple lines of evidence to eliminate or reduce-the salient threats to validity—should be encouraged.

When should one conduct a large-scale RCT? Levy and Sondik (222) describe a complex multiphase, multigroup decision process based on four broad decision criteria: knowledge, methodology, resources, and ethics. Methodological considerations, involving power, significance level, effect size, and the like, are used to estimate the number of subjects and the length of the study. These factors determine the cost of the study and hence its feasibility. In sum, Levy and Sondik outline a complex group decision process that provides a type of cost-benefit analysis for conducting an RCT. The emerging methodology of decision analysis **(386)** would be useful in selecting medical innovations for such high-quality technology assessments.

In conclusion, the dynamic nature of medical innovation requires constant monitoring. This can be accomplished either through postmarketing surveillance (as noted above) or by careful, systematic reviews of the accumulating literature dealing with the innovation. Thus, medical technology assessment must not be viewed as a one-time event. As the model described in table C-2 indicates, evaluative studies for technology assessments should be considered at all stages of development, particularly during the second stage.